View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Electronic archive of Tomsk Polytechnic University

УДК 004.89

РАСШИРЕНИЕ РЕЛЯЦИОННОЙ АЛГЕБРЫ РЕКУРСИВНЫМИ СТРУКТУРАМИ

В.В. Соколова, О.М. Замятина, В.Б. Новосельцев

Томский политехнический университет
E-mail: veronica@tpu.ru

Изложено расширение классической реляционной алгебры, призванное обеспечить возможность отображения составных (иерархических) наборов данных для широкого круга предметных областей. Для достижения этой цели переопределяются основные понятия алгебры Кодда, выделяются существенные для дальнейшего изложения положения реляционной теории и доказывается замкнутость расширенной алгебры.

Ключевые слова:
Реляционная модель данных, база данных, домен, кортеж, рекурсивное отношение.

Key words:
Relational data model, database, domain, tuple, recursive relation.

В настоящее время большинство систем управления базами данных (СУБД) поддерживают реляционный подход, который позволяет отобразить информационную модель предметной области в реляционные отношения [1]. Актуальной темой исследований является отображение предметных областей, содержащих иерархические данные, такие как компании, состоящие из дочерних филиалов, детали, из которых собираются узлы механизмов, в реляционные отношения и их обработка [2]. Для решения данных задач предлагаются различные способы: ввод рекурсивного ключа или дополнительных атрибутов в реляционное отношение [3], обработка иерархий с помощью дополнительных конструкций языка SQL [4]. Каждый из данных способов имеет ограничения в применении. Например, для вывода древовидной структуры с использованием «списка смежных вершин» необходимо точно знать количество уровней вложенности в иерархии. При использовании «метода подмножеств» целостность данных поддерживается триггерами, которые каждый раз перезаписывают список и уровни «предков» узла при его изменении. Метод «вложенного множества» Джо Селко [3] не гарантирует целостность дерева при редактировании его элементов и требует выполнения нескольких запросов для пересчета его левых и правых значений, при этом теряются преимущества «вложенных множеств» для быстрой генерации дерева.

Для сохранения целостности данных, их корректности и неизбыточности предлагается методика отображения иерархических структур в виде вложенных реляционных отношений.

Базовые определения и постановка задачи

Зафиксируем сигнатуру: $\Sigma=<A,D,F>$, где A – множество имен сущностей, D – множество имен доменов, F – множество имен функциональных связей над атрибутами. Все множества сигнатуры являются не более чем счетными. Во множестве D зафиксировано непустое конечное подмножество $T\subset D$ – имен первичных доменов (предопределенных типов). Отметим, что ряд понятий и терминов прямо заимствованы из теории реляционных баз данных и не обсуждаются детально.

Под *модельным объектом* (*модельной сущностью*) понимается формальное в рамках рассматриваемой предметной области определение конкретной сущности, отличимой от другой. Все объекты обладают свойствами, описываемыми *атрибутами*. Под термином *атрибут* будем понимать пару (имя сущности, тип). *Интерпретированный атрибут* – суть подмножество элементов соответствующего типа, *денотативно* выделенное из последнего. Под термином *кортеж* традиционно подразумевается множество пар: атрибут-значение. *Потенциальный ключ* – это минимальный набор атрибутов, однозначно определяющий оставшуюся часть кортежа. *Домен данных* – это (не более, чем счетное) именованное множество некоторых однородных допустимых значений или семантически целостных структур значений. В качестве домена, к примеру, можно рассматривать множество целых.

Домен характеризуется следующими свойствами:
- имеет уникальное имя (в пределах модели данных);
- определен на некотором простом типе данных или на других доменах;
- может иметь некоторые логические свойства, позволяющие специфицировать его (домена) подмножества;
- имеет конкретную смысловую нагрузку.

При задании конкретной интерпретации **I** именам доменов (в том числе базовым) сопоставляются конкретизированные типы (множества). Предполагается, что именам базовых доменов при этом сопоставляются традиционные для программирования типы (например, *integer*, *real*, *char*, *Boolean* и т. д.). Непервичными доменами могут выступать именованные подмножества базовых (например, *Name*$\subset$*string*), либо подмножества декартовых произведений ранее определенных доменов (*Person*$\subset$*Name*$\times$*Age*$\times$*Address*).

Суммируя сказанное, уточним определение домена следующим образом. Пусть задана интерпретация **I**, $P\in T$, $R\in D\setminus T$, тогда:
- $P|_I$ – домен;
- если $R|_I\subset P|_I$, то $R|_I$ – домен;
- если $R_i|_I$ (i=1,..., n) – домены, то $R_0|_I\subset R_1|_I\times...\times R_n|_I$ – домен.

Заметим, что все высказанные ранее соображения относительно семантической целостности доменов остаются актуальными.

Прежде чем переходить к фиксации базового для реляционной алгебры понятия отношения, отметим, что наследники примитивных типов импортируют все операции и отношения от порождающих. Так, для *Age*⊂*Positiv_Integer* имеет место отношение «>» и все операции для целых положительных.

Отношением является некоторое подмножество декартова произведения одного или более доменов: $R(A_1, A_2,..., A_n)$.

Отношение обладает двумя важными свойствами:
- не содержит совпадающих кортежей;
- порядок кортежей несущественен.

Под *реляционной базой данных* будем подразумевать совокупность взаимосвязанных отношений.

Физическим представлением (нерекурсивного) отношения является «плоская» таблица, заголовок которой определяется упорядоченным списком атрибутов, а строки – кортежи – соответствующим образом упорядоченных значений, при этом атрибуты именуют столбцы таблицы. Поэтому иногда говорят «столбец таблицы», имея в виду «интерпретированный атрибут отношения». Этой терминологии придерживаются в большинстве коммерческих реляционных СУБД [4].

Схема домена – это именованное множество пар (имя атрибута, имя домена) вида $D(r)=r.(D_1(a_1),...,D_n(a_n))$, где D, $D_i \in \mathrm{D}(i=1,...,n)$ – имена доменов, r, $a_i \in \mathrm{A}(i=1,..., n)$ – имена атрибутов. В правой части определения r может «проноситься» в скобки на любую глубину, так что $r.(D_1(a_1),...,) \rightleftharpoons (r.D_1(a_1),...,) \rightleftharpoons (D_1(r.a_1),...,)$.

Иногда для большего технического удобства используется индуцируемое схемой таблицы понятие *кортеж-тип*, при этом осуществляется переход от функционального представления к типизированному.

Кортеж-тип для схемы D из предыдущего определения представляется формой $C=<a_1:D_1,a_2:D_2,...,a_n:D_n>$ с аналогичными функциями принадлежности. *Степень* или *арность* схемы отношения определяется количеством атрибутов n схемы.

Реляционная алгебра представляет собой набор операций, использующих отношения в качестве аргументов, и возвращающие отношения в качестве результата. Таким образом, реляционный оператор f выглядит как функция типа «отношения» с отношениями в качестве аргументов: $R=f(r_1,r_2,...,r_n):r_0$. Реляционная алгебра является замкнутой, поэтому в качестве аргументов в реляционные операторы можно подставлять другие реляционные операторы, подходящие по типу: $R=f(f_1(r_{11},r_{12},...,r_{1n}), f_2(r_{21},r_{22},...,r_{2n}),...)$. Таким образом, в реляционных выражениях можно использовать вложенные подвыражения сколь угодной глубины.

Согласно [1], набор операций алгебры определяется восемью дефинициями, которые делятся на два класса: теоретико-множественные и специальные реляционные операции. Приведем определение операции объединения (остальные операции алгебры формулируются аналогично).

Пусть даны два непервичных домена $T_1 \subset D_1$ и $T_2 \subset D_2$, где $D_1 \stackrel{\circ}{=} D_2$ синтаксически равны, тогда результатом операции объединения двух отношений является отношение, включающее все кортежи, входящие хотя бы в одно из отношений-операндов. Таким образом, результат объединения это отношение $T|_{\mathrm{II}}=T_1|_{\mathrm{II}} \cup T_2|_{\mathrm{II}}$, где $(T|_{\mathrm{II}}:(t|_{\mathrm{II}} \in T_1|_{\mathrm{II}}) \vee (t|_{\mathrm{II}} \in T_2|_{\mathrm{II}}))$. Считается, что совпадающие кортежи элиминируются.

Поскольку результатом любой реляционной операции является некоторое отношение, можно образовывать реляционные выражения, в которых вместо отношения-операнда некоторой реляционной операции находится вложенное реляционное выражение.

Рекурсивным называется объект, частично определяемый с помощью самого себя. Ниже в статье допускается только явная рекурсия в смысле следующего определения: отношение $R=(a_1:D_1,a_2:D_2,...,a_n:D_n)$ является *допустимым*, если синтаксическое равенство $D_i \stackrel{\circ}{=} R$, определяющее рекурсию, выполняется не более, чем для $n-1$ доменов. Заметим, что *кортеж* не может ссылаться на самого себя ни непосредственно, ни опосредованно – это *семантическое условие* проверяется (и поддерживается) всеми операциями модификации отношений.

Введем определение частного случая реляционного отношения за счет использования рекурсии. Прежде всего, введем понятие *осно́вного кортежа* (или *основы*): кортеж-тип $C=<a_1:D_1,a_2:D_2,...,a_n:D_n>$, где все атрибуты a_i попарно различны и специфицированы первичными типами будем называть *осно́вным кортежем-типом* или *основой*.

Обобщенным отношением осно́вного отношения R будем называть конечную совокупность регулярных кортежей-значений C_r, определяемых *основным* кортежем-типом C таблицы R, если:
- кортежи-значения C_r попарно различны;
- каждый кортеж $C_r|_{\mathrm{II}}$ есть m-кратно ($m \geq 1$) повторенный набор атрибутов-значений $<a_{11},a_{12},...,a_{1n},<a_{21},a_{22},...,a_{2n},...,<a_{m1},a_{m2},...,a_{mn}...>>>$, таких что, $a_{ik} \in D_k$ $(1 \leq k \leq n)$ основного кортежа C и никакие два подкортежа $<a_{i1},a_{i2},...,a_{in}>$ и $<a_{j1},a_{j2},...,a_{jn}>$ не совпадают при $i \neq j$;
- хотя бы один из атрибутов первого подкортежа имеет значение *nil* (пустое значение) в позиции рекурсивной ссылки (выход из рекурсии).

Для обобщенной таблицы R стандартным образом определены ключевой набор K и индексный набор I:
- $D(K \subset D(I)$ – домен ключей совпадает с доменом индексов;
- $K \cap I=\varnothing$ – никакой ключ не является индексом, что позволяет избежать бесконечной рекурсии;
- определено отображение $M:I \rightarrow K$, причем неподвижной точкой такого транзитивного замыкания M^* является пустое значение *nil*.

Тогда *рекурсивным эквисоединением отношения R* по отношению R называется операция, результатом которой выступает обобщенное отношение R_R, такое что, для каждого её кортежа-значения $C_r|_{\mathrm{II}}$, со-

стоящего более чем из двух кортежей основной таблицы $<c_1,c_2,...,c_m>$:

- $I^+(c_m)=nil$ – ссылочный элемент последнего кортежа является пустым;
- $I^+(c_{k-1})=K(c_k)$, где $k\le m$ – ссылка на непосредственно следующий кортеж определяется функциональной связью.

Итак, мы *расширили реляционную алгебру*, переопределив понятие отношения и соответствующих операций.

Теорема о замкнутости расширенной алгебры

Докажем *теорему о замкнутости расширенной алгебры*: рекурсивные отношения и операции над ними образуют замкнутую систему.

Приведем доказательство теоремы для операции объединения. Формулировка теоремы означает, что результатом применения каждой из введенных операций (объединение, пересечение, вычитание, проекция, селекция, соединение) к произвольным отношениям R_1 и R_2 будет отношение в смысле нового определения.

Теорема. Рассмотрим рекурсивные отношения R_1 и R_2. Пусть $G=R_1\cup R_2$ – отношение, полученное в результате объединения. Докажем, что результирующее отношение G удовлетворяет определению *рекурсивного отношения*, введенному выше.

Доказательство.

1. Поскольку R_1 и R_2 – рекурсивные отношения, то каждая из них состоит из попарно различных кортежей $C'_r(1\le r\le n)$ и $C''_r(1\le r\le m)$ соответственно. Тогда по определению операции объединения отношение G будет также состоять из попарно различных кортежей C_r $(1\le r\le k,\ k\le m+n$ – повторяющиеся кортежи исключаются).
2. Аналогично, в отношении G каждый кортеж есть m-кратно повторенный набор атрибутов-значений $<a_{11},a_{12},...,a_{1n},<a_{21},a_{22},...,a_{2n},...,<a_{m1},a_{m2},...,a_{mk}>>>$, таких, что $a_{is}\in Ds$ $(1\le s\le n)$ основного кортежа C и никакие два подкортежа $<a_{i1},a_{i2},...,a_{ik}>$ и $<a_{j1},a_{j2},...,a_{jk}>$ не совпадают при $i\ne j$.
3. R_1 – отношение в смысле нового определения. Следовательно, хотя бы один из атрибутов первого подкортежа имеет значение *nil*. R_2 – также рекурсивное отношение, поэтому, хотя бы один из атрибутов её первого подкортежа имеет значение *nil*. Тогда отношение G будет также содержать хотя бы один такой подкортеж (по действию операции объединения).
4. R_1 – рекурсивное отношение, следовательно, оно содержит ключевой набор K' и индексный набор I', такие что $D(K')=D(I')$ и $K'\cap I'=\varnothing$. R_2 – также переопределённое отношение, поэтому также содержит ключевой набор K» и индексный набор I'', такие что $D(K'')=D(I'')$ и $K''\cap I''=\varnothing$. Из требования совместимости отношений R_1 и R_2 по объединению и действию операции объединения вытекает, что $D(K)=D(I)$ и $K\cap I=\varnothing$, где $K=K'\cup K''$, $I=I'\cup I''$.
5. R_1 – отношение в смысле нового определения, следовательно, существует отображение $M':I'\to K'$, причем неподвижной точкой такого транзитивного замыкания M'^* является значение *nil*. R_2 – рекурсивное отношение, следовательно, существует отображение $M'':I''\to K''$, причем неподвижной точкой такого транзитивного замыкания M''^* является значение *nil*. В отношении G можно определить отображение $M:I\to K$, которое действует по правилу $M(i)=M'(i)$ при $i\in I'$ и $M(i)=M''(i)$ при $i\in I''$. Транзитивное замыкание M^* будет иметь неподвижные точки, значение которых *nil*.
6. Для каждого кортежа-значения $C_{r|I}$ отношения G, состоящего более чем из двух кортежей основной таблицы $<c_1,c_2,...,c_m>$ будут выполняться условия $I^+(c_m)=nil$ и $I^+(c_{k-1})=K(c_k)$, $k\le m$.

Таким образом, результатом операции объединения является рекурсивное отношение. Аналогичные доказательства несложно осуществить для остальных операций реляционной алгебры.

Сформулированная и доказанная теорема о замкнутости расширенной реляционной алгебры, предоставляет возможность выполнения запросов на обобщенном отношении. Таким образом, предложенный подход может использоваться в качестве теоретического аппарата для построения систем управления базами данных, поддерживающих естественное представление и механизмы манипулирования иерархическими (составными) структурами данных.

Заключение

1. Предложена методика отображения иерархических структур данных в виде вложенных реляционных отношений. Такое представление иерархий повышает информативность модели данных и создает возможности для разработки формальных методов, осуществляющих «свертку» и «развертку» иерархий с любым уровнем вложенности.
2. Доказана теорема о замкнутости расширенной реляционной алгебры на примере операции объединения рекурсивных отношений, что гарантирует обеспечение целостности данных и соответствие формам нормализации.

СПИСОК ЛИТЕРАТУРЫ

1. Дейт К.Дж. Введение в системы баз данных, 7-е изд. – Пер. с англ. – М.: Вильямс, 2002.– 1072 с.
2. Дейт К.Дж., Дарвен Х. Основы будущих систем баз данных. Третий манифест, 2-е изд. – Пер. с англ. – М.: Янус-К, 2004. – 656 с.
3. Celko J. Trees in SQL. Some answers to some common questions about SQL trees and hierarchies // Intelligent enterprise magazine. 2010. URL: http://www.intelligententerprise.com/001020/celko.jhtml?_requestid=1266295 (дата обращения: 01.10.2010).
4. Кайт Т. Oracle для профессионалов. Кн. 1. Архитектура и основные особенности. – Пер. с англ. – М.: DiaSoft, 2003. – 1427 с.

Поступила 07.10.2010 г.